Compound Descriptors in Context: A Matching Function for Classifications and Thesauri

Douglas Tudhope, Ceri Binding,
Dorothee Blocks, Daniel Cunliffe
School of Computing
University of Glamorgan
Pontypridd, CF37 1DL
Wales, UK
Tel. +44 (0)1443 482271
{dstudhope, cbinding, dblocks, djcunlif}@glam.ac.uk

ABSTRACT
There are many advantages for Digital Libraries in indexing with classifications or thesauri, but some current disincentive in the lack of flexible retrieval tools that deal with compound descriptors. This paper discusses a matching function for compound descriptors, or multi-concept subject headings, that does not rely on exact matching but incorporates term expansion via thesaurus semantic relationships to produce ranked results that take account of missing and partially matching terms. The matching function is based on a measure of semantic closeness between terms, which has the potential to help with recall problems. The work reported is part of the ongoing FACET project in collaboration with the National Museum of Science and Industry and its collections database. The architecture of the prototype system and its interface are outlined. The matching problem for compound descriptors is reviewed and the FACET implementation described. Results are discussed from scenarios using the faceted Getty Art and Architecture Thesaurus. We argue that automatic traversal of thesaurus relationships can augment the user's browsing possibilities. The techniques can be applied both to unstructured multi-concept subject headings and potentially to more syntactically structured strings. The notion of a focus term is used by the matching function to model AAT modified descriptors (noun phrases). The relevance of the approach to precoordinated indexing and matching faceted strings is discussed.

Categories and Subject Descriptors
H.3.1 **[Information Storage and Retrieval]**: Thesauruses; H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval; H.3.7 **[Information Storage and Retrieval]**: Digital Libraries; H.5.4 **[Information Interfaces and Presentation]**: Hypertext/Hypermedia

General Terms
Algorithms, Performance, Measurement

Keywords
Compound Descriptors; Term Expansion; Semantic Distance Measures; Similarity Coefficients; Matching Functions; Knowledge Organization Systems; Faceted Classification; Postcoordination; Precoordination

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.

1. INTRODUCTION
The emergence of Digital Libraries has led to renewed interest in the application of Knowledge Organization Systems, such as classifications, gazetteers, ontologies, taxonomies and thesauri to expand free text searching or as controlled vocabulary indexing languages [4, 14, 15, 18]. Such systems model the underlying semantic structure of information and can be used as an aid to indexing, retrieval and resource discovery - "a semantic road map for searchers and indexers" [27]. These semantic relationships provide an opportunity for knowledge-based Digital Library systems, with the system taking an active role in the retrieval process via term suggestion, query expansion and flexible (ranked) matching [e.g. 6, 10, 23, 25, 30]. In particular, a substantial body of work has investigated term expansion based on various measures of semantic distance between thesaurus concepts [e.g. 24, 29]. Underlying many such systems is a matching function that ranks results based upon a similarity coefficient between sets of terms. Here we explore the design issues for a matching function that, given a thesaurus semantic distance measure, yields ranked results for collections where items are indexed by multiple thesaurus terms.

1.1 Definitions
Common to such systems is the notion of a classification based upon hierarchical arrangements of terms (representing concepts). Classifications can be considered as primarily *enumerative*, when all possible simple and compound terms are explicitly listed in their hierarchical position, or as *faceted* [16]. Faceted classifications are based on a primary division of

terminology into fundamental, high-level categories, or *facets*, first named by Kaiser and pioneered by Ranganathan in the analytico-synthetic approach and the Colon Classification. Faceted systems (e.g. BLISS, MESH, PRECIS) are *synthetic*. Rather than attempting to include many possible multi-concept headings or descriptors (and locate them in a hierarchical context), they are synthesised by combination of terms from a limited number of fundamental facets [3, 8]. Thus descriptors are combined as needed at indexing time or at runtime in query formation. Faceted thesauri are similar in structure to faceted classifications but explicitly represent equivalence, hierarchical and associative relationships between concepts [2]. This paper focuses on the J. Paul Getty Trust's Art and Architecture Thesaurus (AAT) [1, 27]. When constructing the AAT, it became apparent that adjectival noun phrases (e.g. *painted oak furniture)* were very common. "Rather than enumerate the nearly infinite number of object and subject descriptions needed by thesaurus users, the AAT decided to pursue the building blocks of these descriptors in the form of a faceted vocabulary" [21 p8]. Descriptors are organised into 7 facets (and 33 hierarchies as subdivisions), representing separate conceptual classes, in an abstract to concrete order: *Associated Concepts, Physical Attributes, Styles & Periods, Agents, Activities, Materials, Objects* (with optional facets for *Time* and *Place*). In fact, the distinction between enumerated and faceted systems is rarely absolute. The major schemes, Dewey Decimal Classification and the Library of Congress Subject Headings (one of the sources for AAT terms), originated as largely enumerative but have become more synthetic as they evolved [9, 16, 23, 28]. The Universal Decimal Classification also has some provision for synthesis.

AAT application guidelines [21] (and precoordinated indexing generally [2, 3, 28]) encourage combination of terms when indexing. AAT descriptors can be *single* concept *descriptors*, *modified descriptors,* or syntactically structured *strings*. Modified descriptors are the most common form of multi-term construction and comprise a *focus* term modified by one or more adjectival terms from other facets (e.g. *Rococo carved gilded wood chairs)* [21 p38]. Strings have more than one focus term and a more complex syntactic structure (e.g. *the renovation of Victorian houses)*. The MARC format field 654 (Subject Added Entry – Faceted Topical Terms) was developed with faceted systems like the AAT in mind. In this paper, we use the term *compound descriptors* to refer generally to AAT modified descriptors and strings. These multi-concept subject headings, built by synthesising single concept base vocabulary elements, allow very specific object descriptions and offer the promise of very precise queries (formed on the same basis). However, practical focus has tended to be on cataloguing rather than searching. The full potential has yet to be exploited in retrieval [5]. The lack of flexible retrieval tools that yield ranked matches of compound descriptors hinders the application of faceted systems. For example, Toni Petersen, then Director of the Art and Architecture Thesaurus Project, outlined key unsolved issues for system designers[1]:

"The major problem lies in developing a system whereby individual parts of subject headings containing multiple AAT terms are broken apart, individually exploded hierarchically, and then reintegrated to answer a query with relevance" [22 p6].

1.2 Matching Function

The issue of a matching function for compound descriptors has seen various approaches by different authors. Previous work [17, 19, 24, 26, 29] has tended to approach the problem as involving unstructured lists of terms and has also identified the issue of retrieval time when incorporating semantic term expansion.

Q: mahogany, dark yellow, gilded, upholstered, armchair

D: oak, light yellow, green, brocade, upholstered,
Carver chair

Figure 1: Query (Q) and compound descriptor (D)

As an illustration of the issues involved, Figure 1 shows an example of a query and compound descriptor discussed later. Note that when matching we need to consider the possibilities of missing, extra, non-matching and partially matching terms.

In the following discussion of the matching function, we are essentially concerned with multi-concept subject headings equivalent to AAT modified descriptors. We return to the issue of strings in the concluding section.

2. THE FACET PROJECT

The work reported here is part of the larger ongoing FACET project investigating the potential of semantic closeness measures in faceted thesauri [12]. This research is in collaboration with the UK National Museum of Science and Industry (NMSI) which includes the National Railway Museum (NRM) [20]. An export of NMSI's collections database (some 400,000 items) is used as the dataset for the research. We have been particularly concerned with the NRM's collections, which have been part of an ongoing project to index areas of 'rich content', in particular the Furnishings and Timekeeping collections. Currently, the AAT is the main thesaurus in the project, although we are also incorporating smaller more specialised thesauri. The AAT is a widely used, large, evolving thesaurus (over 120,000 terms) [13]. The AAT is being piloted in NMSI for specific collections. Coverage spans the Objects, Materials, Physical Attributes and Styles & Periods facets and maps to various database fields, including the Materials and Object name field. As common in museum collections databases, cataloguing includes both index fields and free text descriptions. Relevant index fields are automatically assembled in FACET to form modified descriptors, in which the focus term and mappings from terms to facets are known.

[1] Discussion of the National Art Library database at the Victoria and Albert Museum, UK, with reference to research by David Bearman [5].

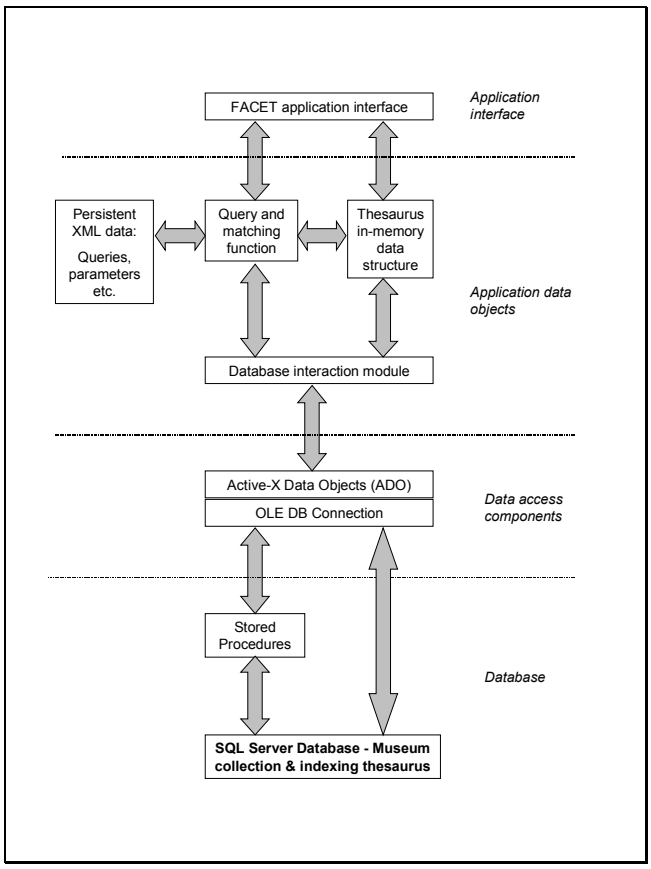

Figure 2: FACET System Architecture

The FACET application is a multi-tiered architecture accessing a SQL Server database, with an OLE DB connection (Figure 2). The thesauri are stored as relational tables in the Server's database. However, a key component of the system is a parallel representation of the underlying semantic network as an in-memory structure of thesaurus concepts (corresponding to preferred terms). The in-memory thesaurus is a directed graph structure utilising the C++ Standard Template Library (STL). The structure models the hierarchical and associative interrelationships of thesaurus concepts via weighted poly-hierarchical links. Its primary purpose is real-time semantic expansion of query terms, achieved by a spreading activation semantic closeness algorithm. There are 28,646 preferred terms in our current AAT version. The in-memory representation yields significant performance benefits for semantic expansion in the matching function, as discussed later. Queries with associated results are stored persistently using XML format data. The intention is to drive the system from an external XML representation of controlling parameters and query structure. The Visual Basic interface (Figure 3) combines different navigable views onto the thesaurus knowledge space on the left, including direct hierarchical context, a thesaurus browser that centres on the selected term and an initial term search facility that takes into account equivalence relationships. The different views are synchronised so that the user can tab from one context to another. The interface includes hypertext navigation history and bookmarking capability for thesaurus terms and colour-coded indicators for facet membership (the icons also indicate the presence of associative links). Any related terms to the currently selected term are shown in the pane at the bottom. Terms are dragged to a direct manipulation Query Builder on the right which maintains the facet structure. The focus term is shown in bold. The underlying Query and Results window is discussed in Section 4.2. We are currently working on interfacing elements of the system to the web and on a 'simple search' version. A module which produces a measure of the semantic closeness between two terms underpins the matching function. This is based on the minimum number of (weighted) transitive relationships that must be traversed in order to connect the terms, with cost factors including depth in the hierarchy. A threshold terminates expansion when terms are considered to be no longer 'close'. The current semantic closeness module is a modified version of the algorithm described in [29], although we are working on a more elaborate version with finer grained control of cost factors.

2.1 Formative Evaluation

An ongoing series of evaluations forms part of the project. In summer 2001, a qualitative formative evaluation of a previous version of the FACET system was conducted with eight museum, library and IT professionals from collaborating institutions [7]. Subjects were assigned tasks and asked to 'think aloud'. Data gathered included transcribed audio recordings, interaction logging, live screen capture and observer notes. The aim of the evaluation was to analyse at a micro level the user's interaction with the prototype in order to reveal problems and inform interface design decisions. Lack of space prevents detailed discussion here, but the evaluation showed that the system, in particular the thesaurus browser and term search facility, was usable. However, occasional window management difficulties resulted from the allocation of search system components to different windows, and users encountered some conceptual problems in faceted query formulation. Problematic issues in result ranking by the algorithm used in the prototype were also identified (discussed below). These findings motivated a revised version of the system with tighter integration of the thesaurus and more support for faceted query formulation and the new matching function described in Section 4, which forms the focus of this paper.

3. MATCHING PROBLEM FOR COMPOUND DESCRIPTORS

A matching function for compound descriptors, which does not rely on an exact match but employs some measure of semantic distance between query-descriptor term pairs, must apply a form of similarity coefficient between the two sets of terms. Figure 1 illustrates the problem. In Rada's pioneering work on medical thesaurus merging and retrieval [24], the Distance metric returned the arithmetic mean of the semantic distances between pairs of terms in query and descriptor (Lee *et al* [19] employ a revised version in their extended Boolean model). The arithmetic mean has the advantage of retaining the metric property but fails to compare like with like (in fact Rada suggests that Distance might be applied selectively when 'semantically distinct dimensions' exist). For example, in Figure

1, the non-match between *armchair* and *oak* would detract from the computed degree of match.

Various authors have demonstrated different versions of a maximal (minimal if expressed as distance) set similarity algorithm which attempts to reduce this problem by taking some average of the *maximum* semantic closeness value achieved by each term in query and descriptor [17, 26] and this was the approach we followed in previous work [29].

While the approach appears intuitive, problems remain in deciding how to calculate the average when faced with missing or non-matching terms in query or descriptor. For example, in Figure 1, should the lack of match for *brocade* in the descriptor penalise the comparison? Problems also arise from treating query and descriptor as unstructured lists of terms. No attempt is made to make use of the information tacitly provided by the facet structure identifying pairs of terms as belonging to the same facet or hierarchy. This problem was highlighted in the above-mentioned evaluation of the first FACET prototype, which re-implemented the earlier maximal set similarity algorithm [29]. Result sets tended to include items not relevant to the main focus of the query due to the effect of auxiliary terms.

Another problematic issue arises from the high computational demand of the matching function since it involves the pair-wise calculation of semantic distances between query and descriptor terms. It is impractical to perform this calculation in real-time on all items in a large collection. Rada applied the Distance function to subsequently rank the results of an initial Boolean query and this is also suggested by Kim & Kim [17]. In their set of experiments on matching functions, Smeaton & Quigley [26] pre-compute term distance values in a global similarity matrix. While pre-computation is a possible approach, we have been interested in real-time computation in order to maintain the potential for dynamically tailoring the semantic closeness measure. However, in spite of restricting matching to a candidate set of information items, our previous work did not achieve a realistic real-time response.

The matching function described in this paper addresses many of the issues raised by Bearman's [5] thought experiment which explores the requirements of a matching function aware of the AAT's facet structure. Bearman notes that there are key unsolved problems for thesaurus search systems that seek to

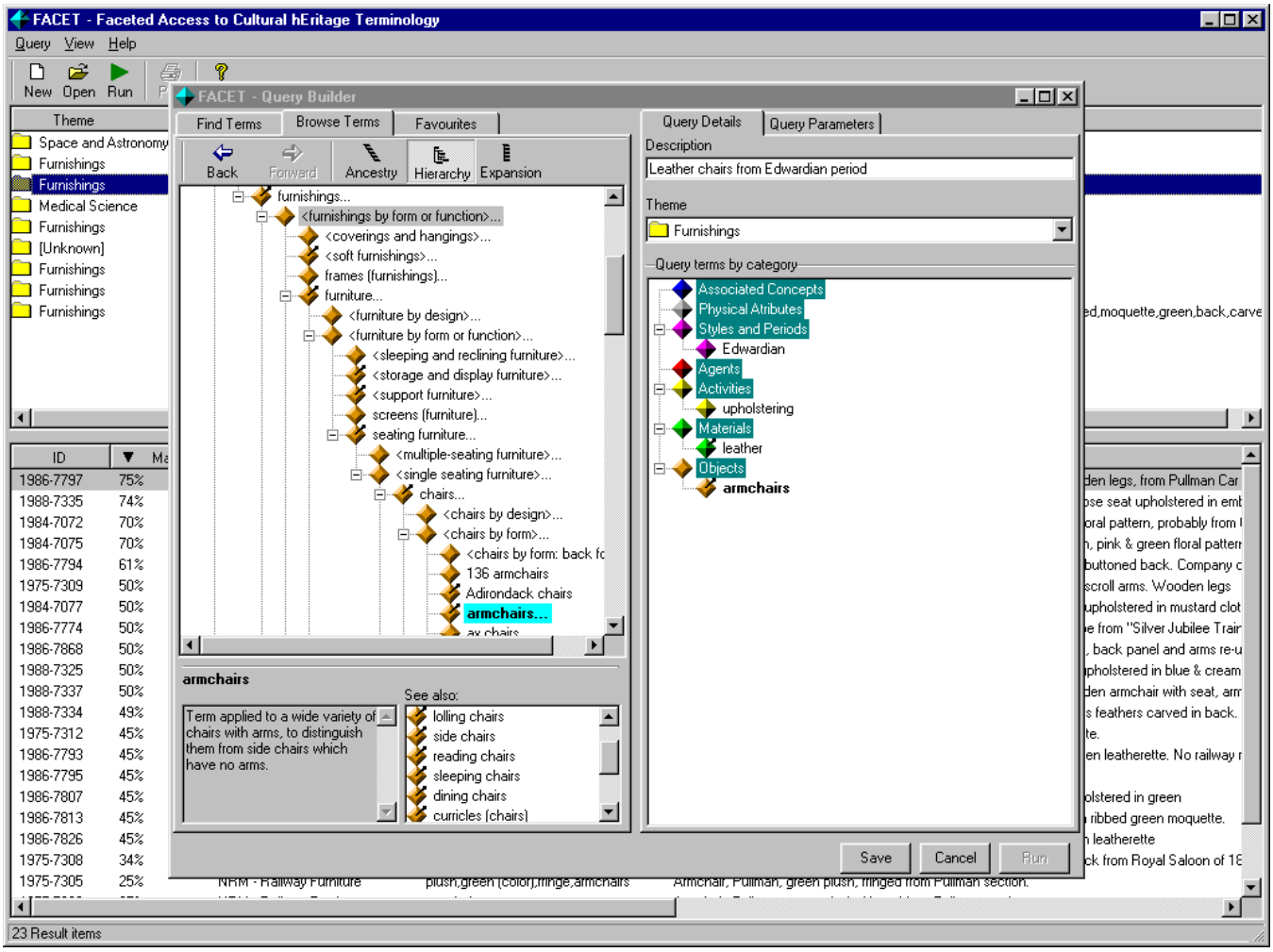

Figure 3: FACET interface showing query builder / thesaurus browser with *armchairs* selected

exploit compound descriptors. He goes on to propose a set of principles for taking advantage of the possibilities allowed in the AAT application protocol. Some of these issues are addressed by FACET's incorporation of semantic closeness into the matching function and by a thesaurus browser that allows a user to interactively refine a search by exploring a term in its hierarchical context. Other principles concern the similarity coefficient. Crucially, Bearman distinguishes between query and collection item descriptor when considering the case of missing (or non-matching) terms. He argues that an extra descriptor term should not result in a penalty. For example, the lack of a Material term in a query should not exclude Objects made of a particular material; there may be differing levels of exhaustivity in collection indexing. However a non-matching descriptor term should be penalised but should not disqualify a possible partial match. In fact, a faceted subject heading, such as *Canadian Victorian painted tables* should be retrieved by queries on broader terms such as *tables*. Bearman is concerned with narrower term expansion but we generalise this to a measure of semantic closeness, which can (selectively) be applied to any of the terms in the query.

4. FACET MATCHING FUNCTION

A similarity coefficient must be able to deal with situations where a match has either extra term(s), missing term(s), non-matching term(s), partially matching term(s), or combinations of the above. The similarity coefficient in the revised matching function was designed to counter the problems encountered in the evaluation described above with the maximal set similarity algorithm. Following Bearman [5], it distinguishes between query and descriptor for non-matching terms and introduces the notion of a focus term. Focus terms are commonly used with respect to faceted thesauri and a similar distinction is sometimes made in automated indexing applications between phrasal heads and modifiers [33]. Here, they serve the dual purpose of focusing the match, to avoid situations described above where subsidiary terms swamp the effect of key terms in the similarity calculation, and they also serve to avoid performance loss due to unnecessary comparisons after semantic expansion. This differs from interactive weighting of query terms since the semantic expansion of the focus term *must* yield a match for an item to be included in results. It also introduces a basic syntactic element to the modified descriptor, which we plan to build on in future string matching work as discussed below. Query focus currently defaults to a term originating from the Objects facet but this can be interactively over-ridden.

4.1 Algorithm

- Perform an expansion using the semantic closeness measure outlined above [29] on the *focus* query term to yield a set of semantically close terms, each with a semantic closeness score. A threshold terminates expansion.
- Find all items in the specified collections, indexed by *at least one* of the expanded set of focus terms. These are the 'candidate items' for a possible match.
- Perform an expansion on the remaining query terms to yield the semantic closeness scores for all non-focus terms.
- For each 'candidate item':

  Get all recognised descriptor terms.

  For each term in the query:

  Establish the semantic closeness score for the closest matching descriptor term.

  Calculate overall degree of match, as in Figure 4d.
- Return a ranked set of items where the overall degree of match exceeds a pre-defined threshold.

4.2 Discussion of Results

Figure 4 compares the operation of the original maximal matching function algorithm (Figure 4b) with the revised algorithm (Figure 4c) for the same query. Results are from indexed records in the NRM collections. The query (Figure 4a) has been chosen to illustrate several design issues. Figure 4d illustrates the similarity coefficient calculation between the query terms (with focus *armchair)* and a candidate item's terms. Note that after semantic closeness expansion, there can be non-matches (0.0), exact matches (1.0) and, importantly, partial matches *(e.g. light yellow – dark yellow)*. The query term *gilded* fails to match, as do descriptor terms *Edwardian,* and *Floral Patterns.* Our original maximal algorithm took the average of both query and descriptor maxima (6.944/13=0.534), penalising richly indexed items. Alternate matches *(light yellow, green)* also degraded the match. The revised algorithm is based on query maxima scores only (3.337/5=0.667), following Bearman's principle of not penalising extra descriptor terms. In fact, the actual formula is slightly more complicated in being a weighted average of each query term's contribution (the default being equal weighting). Thus, for N terms, a non-matching query term penalises by 1/N – a predictable principle for the user. However other weightings could be applied, for example further emphasising the focus term.

A comparison of the top 10 results in Figure 4b (original maximal algorithm) and Figure 4c (revised algorithm) shows that Figure 4b contains item 1978-7574, a *footstool / spittoon,* unlikely to be considered relevant to a query on *armchair* by most users. Note that this item is absent in Figure 4c's result set due to the introduction of the focus term approach. The top item in Figure 4b has an exact match on 3 out of 5 query terms (0.6 score) but drops to second place with the revised algorithm (Figure 4c). The top match in the revised algorithm results shows a narrower term match on armchair and (partially) matches on 4 out of 5 query terms (Figure 4d), which seems a better overall match (0.667 score). Thus we would argue the revised algorithm produces more appropriate results.

Restricting the range of candidate items to those produced by Step 1 and insisting on a focus term match (after expansion) produces performance benefits by restricting the number of similarity comparisons. In this example, the expansion in Step 1 for focus term(s) is Narrower Terms (NT) only, as opposed to non-focus terms where Broader and Related Terms (BT, RT) are also traversed. The rationale is that a user may expect a stricter match on focus terms. However, this may vary with context and could be controllable via a system parameter. The current set of weightings assign a (lesser) cost to NT expansion, in order to differentiate items indexed at the same level of specificity as the search term from items indexed at a more specific level. Due to the bestmatch approach, more specific items appear in the result set but at a lower level. However, the NT cost parameter could be set to zero if desired, which would include all NTs.

mahogany, dark yellow, gilded, upholstered, armchair

Figure 4a: Query terms used for comparison of maximal and revised matching function

Reference	***Match***	***Index terms***
1984-7077	0.600	*armchair, upholstered, dark yellow, wood, cloth*
1986-7774	0.575	*armchair, upholstered, deep yellow, blue*
1984-7072	0.534	*Carver chair, upholstered, mahogany, light yellow, green, pink, Edwardian, floral patterns*
1988-7325	0.471	*armchair, upholstered, light yellow, blue, pattern, wood*
1986-7797	0.427	*armchair, upholstered, green, wood, leather*
1988-7337	0.388	*armchair, upholstered, blue, carved, wood, initials*
1978-7574	0.367	*footstool, spittoon, upholstered, mahogany, brass, blue, grey*
1975-7309	0.362	*armchair, upholstered, blue, buttoned, curved, moquette, scrolled arms, wood*
1986-7868	0.328	*armchair, upholstered, red, carved, imitation leather, inlay, motif, wood*
1975-7305	0.252	*armchair, fringe, green, plush*

Figure 4b: Original maximal algorithm results - top 10 matches

Reference	***Match***	***Index terms***
1984-7072	0.667	*Carver chair, upholstered, mahogany, light yellow, green, pink, Edwardian, floral patterns*
1984-7077	0.600	*armchair, upholstered, dark yellow, wood, cloth*
1988-7325	0.504	*armchair, upholstered, light yellow, blue, pattern, wood*
1986-7774	0.504	*armchair, upholstered, deep yellow, blue*
1988-7337	0.427	*armchair, upholstered, blue, carved, wood, initials*
1986-7868	0.427	*armchair, upholstered, red, carved, imitation leather, inlay, motif, wood*
1986-7797	0.427	*armchair, upholstered, green, wood, leather*
1975-7309	0.427	*armchair, upholstered, blue, buttoned, curved, moquette, scrolled arms, wood*
1988-7335	0.390	*Carver chair, upholstered, brown, embossed, leather, wood, carved, motif, Queen Anne style*
1986-7826	0.390	*Carver chair, upholstered, green, wood, imitation leather*

Figure 4c: Revised FACET algorithm results - top 10 matches

	Query terms					
Index terms	*armchair (focus)*	*gilded*	*upholstered*	*mahogany*	*dark yellow*	***Maxima***
light yellow	0	0	0	0	0.521	**0.521**
Edwardian	0	0	0	0	0	**0**
floral patterns	0	0	0	0	0	**0**
green	0	0	0	0	0.135	**0.135**
upholstered	0	0	1.000	0	0	**1.000**
mahogany	0	0	0	1.000	0	**1.000**
pink	0	0	0	0	0.135	**0.135**
Carver chair	0.816	0	0	0	0	**0.816**
Maxima	**0.816**	**0**	**1.000**	**1.000**	**0.521**	

Figure 4d: Breakdown of item reference 1984-7072 for comparison of calculations

Figure 4: Comparison of maximal and revised matching function on same query

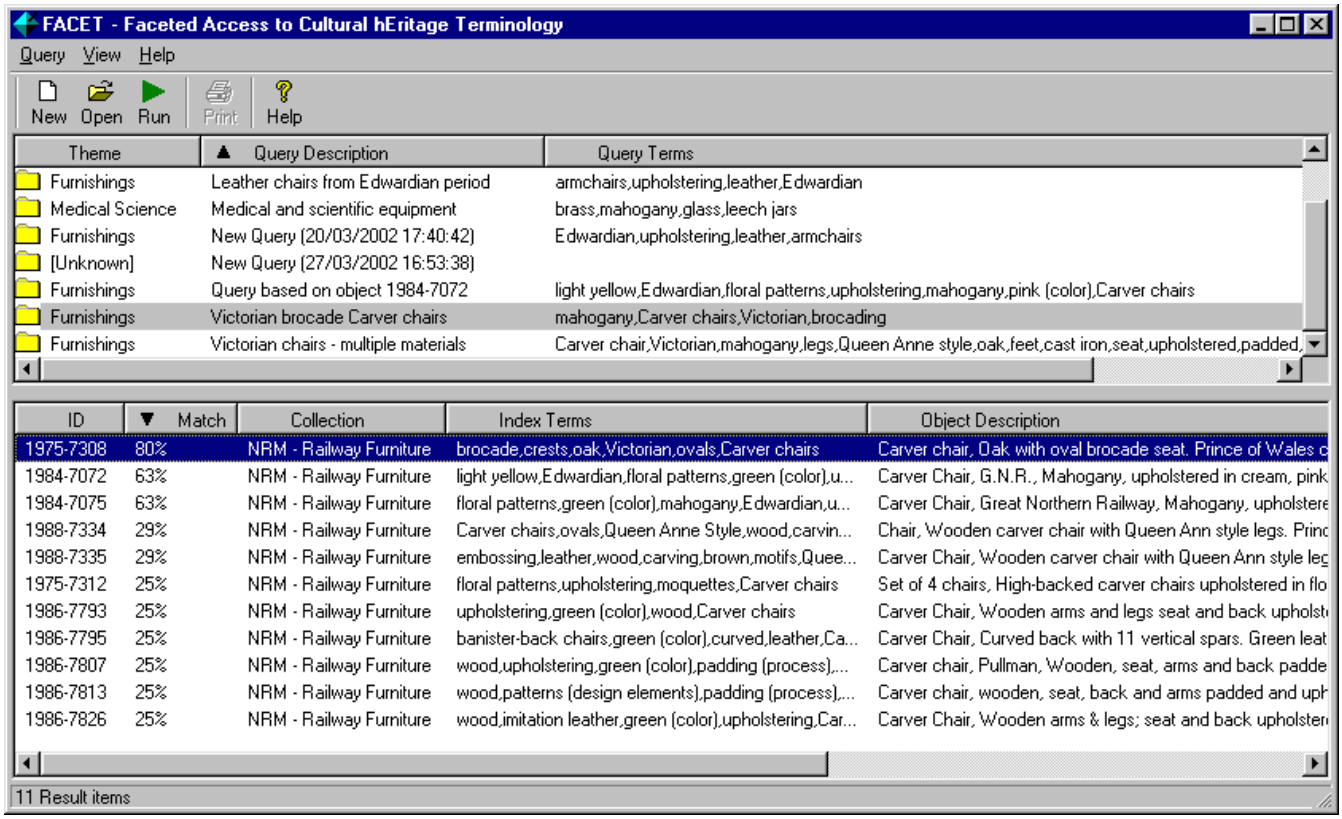

Figure 5: FACET interface showing multiple queries in upper half window, and results from selected query in lower half

Figure 5 shows the interface with several queries in the top half of the window and results for the selected query in the bottom half. A 'theme' is specified as part of the query, indicating a group of sub-collections to be searched within the overall collections database. The selected query shows various full and partial matches on the focus term (*Carver chairs*) and terms from the Materials, Styles & Periods and Activities facets *(Mahogany, Victorian, brocading)*. The top ranked result item (1975-7308 – *brocade, crests, oak, Victorian, ovals, Carver chairs*) illustrates some interesting issues. Descriptors *crests* and *ovals* do not match but do not penalise. The 80% match is derived from two exact matches *(Carver Chairs, Victorian)* and partial matches on *Mahogany-Oak* (both hardwoods) and *brocading-brocade.* The latter pair are in fact terms from different facets *(Activity* and *Material)* connected by an associative relationship in the AAT which yielded the partial match after RT-expansion. In some situations, a user might expect matches for non-focus terms to be subject to a same-facet principle, e.g. for a material to only match with another material after expansion. It would mainly affect extra-facet RT expansion. This facet filter on non-focus term matching would take more account of syntactic role in the descriptor but there is a tension with potential serendipity via RT-expansion, as in this example (see also [31] on RT expansion issues). The cost/benefit depends on query and indexing context and the choice could be parameterised. Although an artificial query scenario, this is an example of the searcher not appreciating the indexing practice for this item. The indexer used a *Material* term to describe the *Object* rather than the *Activity/Process* applied. In this situation, RT-expansion was useful and served to overcome the gap in understanding between indexer and searcher. It also suggests that more formally defining the roles played by facets could be useful – an intelligent Query Builder (or Indexing Editor) might suggest a query structure to follow. We intend to explore facet roles further in future work.

Query	*Maximum*	*Minimum*	*Average*
1	3.485	3.324	3.390
2	0.841	0.360	0.454
3	4.847	3.806	4.498
4	0.751	0.561	0.666
5	1.171	0.892	1.028
6	0.290	0.200	0.227
7	0.351	0.270	0.297
8	0.460	0.351	0.402
Max	4.847	-	-
Min	-	0.200	-
Avg	-	-	1.370

Figure 6: Average query response times (seconds)

Figure 6 displays average response times using the revised algorithm, for repeated execution of eight different queries selected from those shown in Figure 5. The test machine was a Pentium III, 733 MHz PC. The query described in Figure 4 (Query 4 in figure 6) yielded an average response time of approximately 0.7 seconds. The slight variations in the response times observed were due to the underlying operating system running other (unrelated) background processes concurrently. Note that there are many issues involved in scaling

considerations - including the collections searched, the number of terms in a query, the proportion of items indexed with AAT terms (the amount and degree of indexing varies), the number of candidate items, pre-processing, etc. The in-memory structure facilitates efficient term expansion and to date the application appears to give real-time performance using the AAT and NMSI's collections, with investigation continuing.

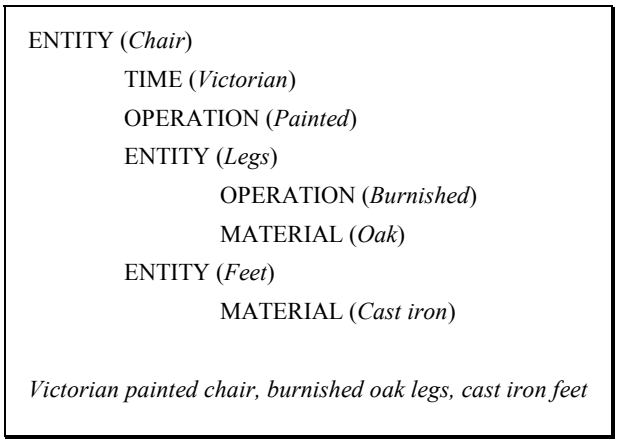

Figure 7: Example of faceted string

5. CONCLUSIONS

This paper has discussed a matching function for compound descriptors that does not rely on exact matching but incorporates term expansion via thesaurus semantic relationships to produce ranked results that take account of missing and partially matching terms. Results suggest that an in-memory semantic network makes a real-time implementation possible. XML representation of query and result sets facilitates interfacing with the wider world. Initial findings from work mapping to a particular facet structure from an external higher level XML representation based on the Classification Research Group's expanded set of categories are reported elsewhere [32].

There are many advantages for Digital Libraries and the web in indexing with Knowledge Organization Systems, whether intellectual or automatic methods are used, but some current disincentive in the lack of flexible retrieval tools that deal with compound descriptors. Semantic term expansion can be applied to a range of hybrid query-navigation techniques, ranging from suggesting possible terms to automatic term expansion and bestmatch query rankings. They may be particularly useful when multi-concept headings are involved and the user is faced with several dimensions of hierarchical context. It may not be practical for a user to browse several hierarchical dimensions and try numerous combinations to exactly match all item descriptors that might be considered relevant, taking into account both indexing exhaustivity (number of terms) and specificity (level of detail). Automatic traversal of relationships can augment the user's browsing possibilities. The techniques can be applied both to unstructured multi-concept subject headings and potentially to more syntactically structured strings.

5.1 Precoordinated Indexing and Strings

The matching function described here essentially operates on modified descriptors (in AAT terminology). In future work we intend to extend the matching to strings with more than one focus term, making use of faceted syntactical structure. For example, Figure 7 shows an item descriptor from a prototype structured query builder. The hierarchical component parts would allow a future matching function to make use of the information that the legs are made of oak while the feet are made of cast iron. The legs are burnished but the chair is painted. This is an object-oriented museum collection example but precoordinated syntactic structure also occurs in multi-concept subject headings (e.g. *brick churches restoration masons)*.

In postcoordinated indexing, several single-concept terms are applied to an item and each can be an access point or they can be combined in a Boolean search. However this can result in 'false drops' – false syntactic associations (e.g. cast iron legs in Figure 7). Precoordination resolves potential ambiguities in the syntactical association of terms. Precoordinated indexing [21, 28] is seen to offer advantages for specificity because terms are placed within a particular context or syntax and this can lead to potential gains in precision on retrieval. However there are also potential problems for recall. A searcher may find it difficult to generate the same string of terms as the indexer and in the same sequence. PRECIS, probably the most highly developed precoordinated indexing system to be used operationally, employed an elaborate rule set over its coded syntactic operators to automatically generate multiple permutations of index string term order to facilitate alphabetic lookup in a printed subject index [3]. This is not necessary in digital environments. For example, Dykstra advocates postcoordinate searching on component terms for online versions of PRECIS [11]. She describes Boolean search statements with the Canada National Film Board's PRECIS online system and suggests the possibility of replacing or enhancing Boolean search to take account of the syntactic role of a term [11]. With Boolean searching, the searcher may still be required to articulate the same combination of terms at the same level of specificity as the indexer. An extended faceted matching function would help with recall problems caused by missing terms and partial matches (via semantic term expansion) and would address some of the issues raised by Dykstra. It would also offer bestmatch advantages over Boolean search in the generation of ranked results.

5.2 Future Work

We intend to revisit in finer grained detail the semantic closeness measure, the weighting of relationships and whether additional cost factors should be incorporated into the expansion. The XML representation of thesaurus/collection mappings and query structure will be developed so that controlling parameters and mappings can be held externally. Further evaluation will be conducted on the revised interface and matching function. Automatically deconstructing strings of terms is another area of future research, applicable both to free text queries and to manual and automatic indexing tools.

6. ACKNOWLEDGEMENTS

We would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (Grant GR/M66233/01) and the support of HEFCW for the Internet Technologies Research Lab. We would like to thank the anonymous reviewers for their helpful comments, Helen Ashby, Ann Borda, Sarah Norville, Charlotte Stone and other staff from the National Museum of Science and Industry for their assistance and the J. Paul Getty Trust for provision of the AAT.

7. REFERENCES

[1] Art and Architecture Thesaurus. J. Paul Getty Trust. http://www.getty.edu/research/tools/vocabulary/aat/

[2] Aitchison J., Gilchrist A., Bawden D. 2000. Thesaurus construction and use: a practical manual (4th edition). London: ASLIB.

[3] Austin D. 1984. PRECIS: a manual of concept analysis and subject indexing. London: British Library.

[4] Batty D. 1998. WWW – Wealth, weariness or waste: Controlled vocabulary and thesauri in support of online information access. D-Lib Magazine, November http://www.dlib.org/dlib/november98/11batty.html

[5] Bearman D. 1994. Thesaurally mediated retrieval. Visual Resources, Vol. 10, 295-307.

[6] Beaulieu M. 1997. Experiments on interfaces to support query expansion. Journal of Documentation, 53(1), 8-19.

[7] Blocks D., Binding C., Cunliffe D., Tudhope D. 2002. Evaluation of information seeking using thesauri in the context of museum collection systems. Technical Report CS-02-1, School of Computing, University of Glamorgan, Pontypridd, CF37 1DL, UK.

[8] Broughton V. 2001. Faceted classification as a basis for knowledge organization in a digital environment: The BLISS Bibliographic Classification as a model for vocabulary management and the creation of multi-dimensional knowledge structures. New Review of Hypermedia and Multimedia, Vol. 7, 67-102.

[9] Chan L., Childress E., Dean R., O'Neill E., Vizine-Goetz D. 2001. A faceted approach to subject data in the Dublin Core metadata record. Journal of Internet Cataloging, 4(1-2), 35-47.

[10] Chen H., Ng T., Martinez J., Schatz B. 1997. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the Worm Community System. Journal of the American Society for Information Science, 48(1), 17-31.

[11] Dykstra M. 1989. PRECIS in the online catalog. Cataloguing and Classification Quarterly, 10(1-2), 81-94.

[12] FACET Research Project. University of Glamorgan. http://web.glam.ac.uk/schools/soc/research/hypermedia/facet_proj/index.php

[13] Harpring P. 1999. How forcible are the right words: overview of applications and interfaces incorporating the Getty vocabularies. Proc. Museums and the Web 1999. Archives and Museum Informatics. http://www.archimuse.com/mw99/papers/harpring/harpring.html

[14] Hill L. 2000. Core elements of digital gazetteers: placenames, categories, and footprints. Proc. 4th European Conference on Research and Advanced Technology for Digital Libraries (J. Borbinha, T. Baker eds.). Lecture Notes in Computer Science, Berlin: Springer, 280-290.

[15] Hodge G. 2000. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. The Digital Library Federation Council on Library and Information Resources. http://www.clir.org/pubs/abstract/pub91abst.html

[16] Iyer. H. 1995. Classificatory structures: concepts, relations and representation. Frankfurt: INDEKS Verlag.

[17] Kim Y., Kim J. 1990. A model of knowledge based information retrieval with hierarchical concept graph. Journal of Documentation, 46(2), 113-136.

[18] Koch T. 2000. Quality-controlled subject gateways: definitions, typologies, empirical overview. Online Information Review, 24(1), 24-34.

[19] Lee J., Kim H., Lee Y. 1993. Information retrieval based on conceptual distance in ISA hierarchies. Journal of Documentation, 49(2), 113-136.

[20] National Museum of Science and Industry (NMSI). http://www.nmsi.ac.uk

[21] Petersen P., Barnett P. (Eds.) 1994. Guide to indexing and cataloging with the Art & Architecture Thesaurus. Oxford: OUP.

[22] Petersen T. 1994. The National Art Library and the AAT (Part II). Art and Architecture Thesaurus Bulletin 22, 6-8.

[23] Pollitt, A. 1998. The application of Dewey Classification in a view-based searching OPAC. Proc. 5th International ISKO conference, Lille, Ergon Verlag: 176-183.

[24] Rada R., Mili H., Bicknell E., Blettner M. 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1), 17-30.

[25] Schatz B., Johnson H., Cochrane P. 1996. Interactive term suggestion for users of digital libraries: using subject thesauri and co-occurrence lists for information retrieval. Proc. 1st ACM International Conference on Digital Libraries, 126-133.

[26] Smeaton A., Quigley I. 1996. Experiments on using semantic distances between words in image caption retrieval. Proc. 19th ACM SIGIR Conference, 174-180.

[27] Soergel. D 1995. The Art and Architecture Thesaurus (AAT): a critical appraisal. Visual Resources, 10(4), 369-400.

[28] Svenonius E. 2000. The intellectual foundation of information organization. Cambridge, MA: MIT Press.

[29] Tudhope D., Taylor C. 1997. Navigation via similarity: automatic linking based on semantic closeness. Information Processing and Management, 33(2), 233-242. Elsevier Science.

[30] Tudhope D., Cunliffe D. 1999. Semantically indexed hypermedia: linking information disciplines. ACM Computing Surveys, Electronic Symposium on Hypertext and Hypermedia, 31(4es).

[31] Tudhope D., Alani H., Jones C. 2001. Augmenting thesaurus relationships: possibilities for retrieval. Journal of Digital Information, 1(8), http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Tudhope/

[32] Tudhope D., Binding C., Blocks D., Cunliffe D. 2002. Representation and retrieval of faceted systems. Proc. 7th International Society of Knowledge Organization Conference (ISKO 2002), Granada, forthcoming.

[33] Wacholder N., Evans D., Klavans J. 2001. Automatic identification and organization of index terms for interactive browsing. Proc. 1st ACM/IEEE-CS Joint Conference on Digital Libraries, 126-134.